BAB 2

TINJAUAN PUSTAKA

2.1 Kinerja Akademik Mahasiswa

Kinerja akademik merupakan hasil akhir yang dicapai oleh siswa atau mahasiswa sebagai keberhasilan selama mengikuti pendidikan dalam sebuah institusi pendidikan (Naomi and Nindyati, 2008). Tingkat pencapaian yang diperoleh mahasiswa melalui akademik yang tidak hanya sekedar diukur dari mengikuti perkuliahan, presentase kehadiran, penyelesaian tugas kuliah, dan ikut aktif dalam kegiatan akademik lainnya, tetapi keberhasilan mahasiswa dalam bidang akademik juga ditandai dengan prestasi akademik yang dicapai dengan ditunjukkan melalui Indeks Prestasi (IP) maupun indeks Prestasi Kumulatif serta ketepatan dalam menyelesaikan studi (Dian Indriana TL, Amerti Irvin Widowati, 2017).

Tingkat kelulusan adalah kriteria utama untuk mengukur kinerja lembaga akademik mana pun sehingga institusi pendidikan terus berusaha memajukan layanan akademik untuk menghadapi kegagalan mahasisiwa (Rajendran, Chamundeswari and Sinha, 2022). Kinerja mahasiswa menjadi bagian penting dalam lembaga pendidikan tinggi sebagai salah satu kriteria perguruan tinggi yang berkualitas berdasarkan catatan prestasi akademiknya yang sangat baik (Shahiri, Husain and Nur'Aini Abdul Rashid, 2015). Oleh karena itu, diperlukan parameter untuk mengukur faktor yang mempengaruhi kesuksekan akademik mahasiswa dan

mengidentifikasi faktor kunci yang mempengaruhi kinerja siswa dalam mencapai target kelulusan yang telah ditetapkan oleh institusi.

Informasi penting berkaitan dengan faktor yang dapat mempengaruhi kesuksesan akademik memiliki dampak yang penting terhadap pembejaran dan pendidikan terutama dalam pencapaian prestasi mahaisiswa (Naomi and Nindyati, 2008). Berbagai penelitian pendidikan telah berusaha mengungkap berbagai faktor yang dapat mempengaruhi tingkat pencapaian akademik mahasiswa. Sebagian besar data demografis dan data akademik digunakan untuk prediksi kinerja siswa (Sathe and Adamuthe, 2021).

2.1.1 Data Demografis Mahasiswa

Faktor-faktor demografis mahasiswa dalam penelitian ini meliputi latar belakang pendidikan sebelumnya, gender, usia masuk, jarak tempuh atau tempat tinggal dan keadaan sosial ekonomi orang tua.

1. Latar belakang pendidikan sebelumnya

Tingkat masuknya siswa ke sekolah memainkan peran penting dalam menentukan keberhasilan di masa depan (Kikas *et al.*, 2009). Latar belakang pendidikan sebelumnya memungkinkan dapat meningkatkan peluang siswa untuk mendaftar ke universitas maupun meningkatkan kinerja siswa setelah masuk dan berada di universitas (Crawford, 2014).

Selanjutnya, (Crawford, 2014) juga menambahkan bahwa terdapat perbedaan potensial antara kinerja siswa dari sekolah swasta dengan sekolah Negeri setelah masuk di Universitas, dimana rata-rata siswa dari sekolah independen cenderung

berkinerja lebih buruk di universitas daripada siswa sekolah negeri (Crawford, 2014).

2. Jenis kelamin

Jenis kelamin atau Gender bersama beberapa variabel demografis lainnya merupakan sebagai penentu dari prestasi akademik (Oluwagbenga Abiodun and Isaiah, 2015), Prediktor Gender menjadi prediktor yang sigifikan dalam menentukan prestasi akademik siswa (Yazici, Seyis and Altun, 2011). Sebagian besar penelitian telah mengidentifikasi adanya perbedaan gender dalam pencapaian pendidikan yaitu rata-rata perempuan cenderung memiliki tingkat pendidikan dan prestasi yang lebih tinggi daripada laki-laki, dimana keunggulan perempuan dalam pencapaian pendidikan terwujud dalam berbagai dimensi dan dari sekolah awal hingga perguruan tinggi (Delaney and Devereux, 2021).

Perempuan Dalam pendidikan tinggi sering ditemukan mengungguli laki-laki terutama kesuksesan dalam hal nilai yang lebih tinggi, karena perempuan cenderung bekerja lebih teliti dan memiliki etos kerja yang lebih kuat seperti sering menghadiri kelas daripada laki-laki (Dayioğlu and Türüt-Aşik, 2007).

3. Usia masuk

Usia memerankan fungsinya dalam memperoleh ketrampilan kecerdasan emosional dan status intelektualnya yang mengarah pada kesiapan untuk sekolah, prestasi akademik, dan produktivitas (Waliyi Olayemi and Olayemi, 2018). Prinsip dasar emosional juga berfungsi sebagai prinsip dasar motivasi yang mendorong anak untuk belajar (Momanyi, Too and Simiyu, 2015). Seiring bertambahnya usia biasanya mempengaruhi berbagai perubahan perkembangan

kinerja manusia, dimana siswa yang lebih tua lebih termotivasi dan lebih berpengalaman dalam banyak bidang kehidupan seharusnya memperoleh nilai rata-rata yang lebih tinggi (Wambugu and Emeke, 2019).

Pada beberapa penelitian yang telah dilakukan oleh beberapa peneliti menemukan bahwa perbedaan usia memiliki pengaruh yang signifikan terhadap prestasi akademik, dimana siswa yang lebih tua mencapai prestasi akademiknya yang lebih tinggi daripada siswa yang lebih muda baik di sekolah menengah mupun memasuki perguruan tinggi (Nam, 2014). Terdapat juga hasil penelitian yang menemukan bahwa menunda masuk sekolah memiliki efek positif pada IPK, kehadiran dan kemungkinan kelulusan (Cáceres-Delpiano and Giolito, 2018).

4. Jarak Tempat Tinggal

Jarak tempuh atau lokasi menjadi salah satu faktor yang dapat mempengaruhi hasil belajar. Beberapa penelitian menemukan hasil bahwa jarak tempat tinggal berkiatan dengan jarak tempuh yang harus dilalui oleh siswa dapat menyebabkan kelelahan secara fisik dan psikologis yang berdampak pada konsentrasi belajar yang buruk di dalam kelas setelah menghabiskan waktu berjam-jam di jalan untuk sampai ke tempat belajar dan perjalanan pulang kembali (Baliyan and Khama, 2020).

Tingkat konsentrasi yang rendah selama proses belajar akan mengurangi kualitas belajar peserta didik dalam memperoleh pengalaman langsung, mengamati sendiri, meneliti sendiri, menyusun dan menyimpulkan sendiri pengetahuan yang diperoleh di dalam kelas sehingga akan menurunkan prestasi belajarnya (Andhika, Floristia and Alawiyah, 2020). Jarak jauh yang ditempuh

oleh siswa akan mengurangi waktu kontak dengan tenaga pendidik, menurunkan tingkat kehadiran dan penyelesaian tugas sekolah (Oneya and Onyango, 2021).

Jarak atau lokasi tempat tinggal yang jauh dari kampus juga akan memakan waktu yang lebih banyak diperjalanan yang mengakibatkan kelelahan akan mempengaruhi kemampuan syaraf sehingga berimbas pada penurunan tingkat konsentrasi belajar yang berdampak pada hasil belajar yang dicapai (Andhika, Floristia and Alawiyah, 2020). Terdapat hubungan negatif antara jarak dengan IPK yang diartikan bahwa dengan meningkatnya jarak perjalanan yang ditempuh siswa berdampak pada kecenderungan penurunan nilai IPK (Nelson *et al.*, 2016).

5. Status sosial ekonomi orang tua

Banyak variabel dalam latar belakang keluarga memiliki hubungan yang kuat (langsung dan tidak langsung) dengan suksesan siswa di seluruh sekolah dan dalam pencapaian pendidikan seperti Variabel struktur keluarga (status sosial ekonomi dan keluarga utuh/orang tua tunggal), tingkat pendidikan orang tua, keterlibatan orang tua dan gaya pengasuhan orang tua (Jacobs and Harvey, 2005).

Keberhasilan suatu pendidkan menjadi salah satu tanggungjawab dari keluarga terutama orang tua dalam memberikan pengarahan dan bimbingan. Keluarga memengaruhi perilaku belajar dan prestasi akademik siswa dengan cara yang penting, karena keluarga merupakan lingkungan utama dan paling signifikan yang dihadapi siswa (Li and Qiu, 2018).

a. Pendapatan orang tua

Dukungan keluarga seperti status sosial ekonomi orang tua menjadi salah satu yang mempengaruhi prestasi akademik mahasiswa (Sirin, 2005); (Cheadle, 2008).

Status sosial ekonomi tidak hanya mencakup pendapatan tetapi juga pencapaian pendidikan, keamanan finansial, dan persepsi subjektif dari status sosial dan kelas sosial (American Pyschological Association, 2017); (Nunes *et al.*, 2023); (Rajendran, Chamundeswari and Sinha, 2022); (Shahiri, Husain and Nur'Aini Abdul Rashid, 2015). Salah satu fitur sosial ekonomi orang tua mencakup potensi keuangan dan elemen sosial budaya. Pendapatan keluarga memiliki pengaruh penting baik langsung maupun tidak langsung terhadap keberhasilan akademik seorang siswa(Tomul and Savasci, 2012).

Adanya perbedaan siginfikan antara status sosial ekonomi orang tua dengan prestasi akademik siswa yaitu dengan tingkat sosial ekonomi tinggi maka prestasi akademik siswa juga tinggi (Yazici, Seyis and Altun, 2011), (Kormos and Kiddle, 2013). Kondisi ini tentunya bukan tanpa alasan, dimana orang tua dengan status sosial ekonomi yang lebih tinggi mempersiapkan anak-anak mereka untuk sekolah secara lebih memadai daripada mereka yang berasal dari kelompok bawah (Jackson Nyamubi, 2019). Sementara Orang tua dengan status sosial ekonomi yang lebih rendah dan ketidakmampuan belajar atau hasil psikologis negatif lainnya yang memengaruhi prestasi akademik seperti kurangnya perhatian, ketidaktertarikan, dan kurangnya kerja sama anaknya di sekolah, persepsi stres ekonomi keluarga dan kendala keuangan pribadi mempengaruhi tekanan/depresi emosional pada siswa dan hasil akademik mereka (American Pyschological Association, 2017).

b. Pendidikan orang Tua

Fitur sosial ekonomi selanjutnya pada lingkup keluarga adalah tingkat pendidikan orang tua. Tingkat pendidikan orang tua merupakan prediktor kinerja siswa dengan berbekal kemampuan orang tua untuk menyediakan lingkungan belajar yang mendukung (Jackson Nyamubi, 2019). Oang tua yang memiliki tingkat pendidikan yang lebih tinggi akan lebih siap membantu anak-anak mereka berkembang secara intelektual dengan mengajak berdiskusi tentang pendidikan, kebiasaan membaca dan pemberian pengembangan keterampilan menginterpretasi dan berfikir kritis (Nunes *et al.*, 2023).

Dengan adanya keterlibatan orang tua tidak hanya akan memengaruhi pembelajaran siswa saat ini, tetapi juga akan meletakkan dasar yang baik dan meningkatkan pembelajaran di masa depan (Keith *et al.*, 1998). Oleh karena itu, Status pendidikan orang tua merupakan salah satu variabel penting yang menjelaskan tingkat pendidikan orang tua memiliki pengaruh langsung dan tidak langsung terhadap prestasi akademik seorang anak terutama Keluarga yang memiliki tingkat pendidikan yang lebih tinggi memiliki kesempatan untuk menyediakan sumber daya ekonomi dan sosial yang lebih mudah dan lebih besar yang akan berkontribusi pada pencapaian akademik anak-anaknya(Tomul and Savasci, 2012)

2.1.2 Data Akademik

Data akademik mahasiswa seperti nilai mata kuliah atau nilai ujian sebelumnya (Sathe and Adamuthe, 2021). Nilai akademik dapat digunakan sebagai indikator kecerdasan mahasiswa. Kecerdasan adalah prediktor tunggal terbaik dalam

mengukur hubungan kecerdasan dengan kinerja akademik siswa (Kuncel, Hezlett and Ones, 2004).

Kecerdasan personal mahasisiwa diukur menggunakan kemampuan kognitif dengan nilai kumulatif yang diperoleh pada tahun pertama saat kuliah (Kuncel, Ones and Sackett, 2010). Mengingat kinerja akademik mahasiswa Kecenderungan menunjukkan bahwa sebagian besar kinerja yang buruk dihasilkan dari kegagalan tahap awal terutama pada tahun pertama pendidikan (Ghosh and Janan, 2021).

Melalui nilai komulatif setiap semester dapat digunakan sebagai rekam jejak akademik seperti kemajuan dan hasil akademik yang telah mahasiswa capai selama masa studi.

2.2 Data Mining

2.2.1 Pengertian Data Mining

Data mining disebut juga dengan *Knowledge Discovery In Database* (KKD) merupakan analisis dari kumpulan data pengamatan untuk menemukan hubungan yang tidak terduga serta untuk meringkas data dengan cara baru yang dapat dipahami dan berguna bagi pemilik data (Badr *et al.*, 2016). Proses *Knowledge Discovery In Database* (KKD) menggabungkan matematika yang digunakan untuk menemukan pola menarik dalam data dengan seluruh proses penggalian data dan menggunakan model yang dihasilkan untuk diterapkan ke kumpulan data lain guna memanfaatkan informasi untuk beberapa tujuan (Robert Nisbet, Gary Miner, 2018). Data mining merupakan bagian integral dari penemuan pengetahuan dalam database (KDD) yang merupakan keseluruhan

proses mengubah data mentah menjadi informasi yang berguna melalui proses yang terdiri dari serangkaian langkah transformasi mulai dari *preprocessing* data sampai *postprocessing* hasil data mining (Tan *et al.*, 2014).

Data mining adalah proses menemukan pola dan tren yang berguna dalam kumpulan data yang besar (Larose and Larose, 2015). Hasil data mining sebagai langkah penting untuk menemukan pengetahuan dalam proses database dengan mengekstrak pengetahun tersembunyi seperti pola, hubungan atau aturan dari kumpulan data besar sehingga dapat dianalisis dan mampu digunakan untuk memprediksi tren masa depan (Mienye, Sun and Wang, 2019). Secara garis besar metode pelatihan yang digunakan dalam teknik-teknik *data mining* dibedakan ke dalam dua (Bendesa Subawa, 2019) yaitu :

1) *Unsupervised learning*

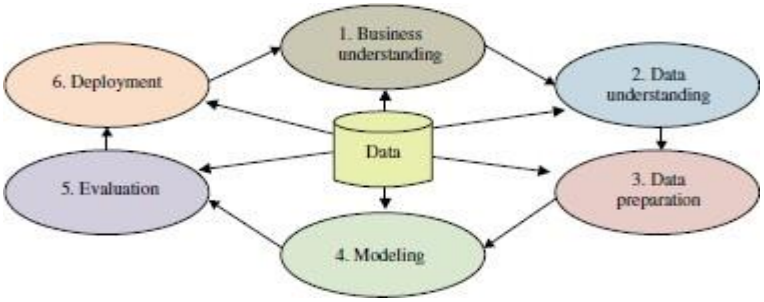
Metode ini diterapkan tanpa adanya latihan (*training*) dan tanpa ada guru (*teacher*). Guru di sini adalah label dari data. Algoritme *data mining* mencari pola dari semua variabel (atribut). Variabel (atribut) yang menjadi target/label/*class* tidak ditentukan (tidak ada).

2) *Supervised learning*

Metode belajar dengan adanya latihan dan pelatih. Dalam pendekatan ini, untuk menemukan fungsi keputusan, fungsi pemisah atau fungsi regresi, digunakan beberapa contoh data yang mempunyai *output* atau label selama proses *training*. Sebagian besar algoritme *data mining (estimation, prediction/forecasting,* klasifikasi) adalah *supervised learning.* Variabel yang menjadi target/label/*class*

sudah ditentukan dan algoritme melakukan proses belajar berdasarkan nilai dari variabel target yang terasosiasi dengan nilai dari *variable predictor*.

Pendekatan terstruktur dalam merencanakan proyek penambangan data untuk mengekspresikan proses analisis/penambangan data prediktif yang tersedia adalah paling lengkap Model konseptual CRISP-DM (Robert Nisbet, Gary Miner, 2018). CRISP–DM merupakan Proses Standar Lintas Industri untuk Data Mining yang menuntut agar data mining dilihat sebagai keseluruhan proses, mulai dari komunikasi masalah bisnis, melalui pengumpulan dan pengelolaan data, pra-pemrosesan data, pembuatan model, evaluasi model, dan penerapan model (Larose, 2005). Tahapan proses CRISP-DM dapat dilihat pada gambar 2.1:



Gambar 2.1 Tahapan proses CRISP-DM (Kantardzic, 2020)

1) *Business understanding*

Pada tahap ini, sebelum memulai penambangan data maka harus dilakukan adalah yang Pertama, nyatakan dengan jelas tujuan dan persyaratan proyek dalam kaitannya dengan bisnis atau unit penelitian secara keseluruhan, Kemudian, terjemahkan tujuan dan batasan ini ke dalam rumusan definisi masalah data mining dan Terakhir, siapkan strategi awal untuk mencapai tujuan tersebut.

2) *Data understanding*

Pada Fase pemahaman data yang dilakukan adalah mengumpulkan data Kemudian gunakan analisis data eksplorasi untuk menemukan wawasan awal, mengevaluasi kualitas data dan yang terkhir memilih himpunan bagian menarik yang mungkin berisi pola yang dapat ditindaklanjuti.

3) *Data Preparation*

Pada Fase ini mencakup semua aspek penyiapan kumpulan data akhir yang akan digunakan untuk fase selanjutnya mulai dari data awal, mentah, dan kotor. Kemudian memilih kasus dan variabel yang ingin dianalisis dan yang sesuai untuk dianalisis, kemudian melakukan transformasi pada variabel tertentu, jika diperlukan, membersihkan data mentah sehingga siap untuk alat pemodelan

4) *Modeling*

Pada Fase pemodelan, langkah yang dilakukan adalah memilih dan menerapkan teknik pemodelan yang sesuai, kalibrasi pengaturan model untuk mengoptimalkan hasil, beberapa teknik yang berbeda dapat diterapkan untuk masalah data mining yang sama, mungkin memerlukan pengulangan kembali ke fase persiapan data, untuk membawa bentuk data sesuai dengan persyaratan khusus dari teknik penambangan data tertentu.

5) *Evaluation*

Pada Tahap Evaluasi, langkah yang dilakukan adalah pada fase pemodelan telah menghasilkan satu atau lebih model. Model-model ini harus dievaluasi kualitas dan efektivitasnya sebelum diterapkan dan digunakan di lapangan serta tentukan apakah model itu benar-benar mencapai tujuan yang ditetapkan, beberapa aspek

penting dari masalah bisnis atau penelitian belum cukup diperhitungkan, Akhirnya, sampai pada keputusan mengenai penggunaan hasil data mining.

6) *Deployment Phase*

Pembuatan model tidak menandakan selesainya proyek. Perlu memanfaatkan model yang dibuat.

Contoh penerapan sederhana: Buat laporan, Contoh penerapan yang lebih kompleks: Menerapkan proses penambangan data paralel di departemen lain, Untuk bisnis, pelanggan sering melakukan penerapan berdasarkan model Anda.

2.2.2 Teknik data mining

Teknik yang dimiliki data mining secara umum berdasarkan tugas yang dapat dilakukan (Larose, 2005) yaitu antara lain:

1) *Description*

Yaitu teknik yang Mencoba menemukan cara untuk menggambarkan pola dan tren tersembunyi dalam data. Deskripsi berkualitas tinggi seringkali dapat dicapai dengan analisis data eksplorasi, metode grafis untuk mengeksplorasi data untuk mencari pola dan tren.

2) *Estimation*

Yaitu teknik yang mirip dengan klasifikasi kecuali variabel tujuan lebih numerik. Dalam estimasi, diperkirakan nilai variabel target numerik menggunakan satu set variabel prediktor numerik dan/atau kategoris. Model dibangun menggunakan catatan "lengkap", yang memberikan nilai variabel target serta prediktor Kemudian, untuk observasi baru, dibuat estimasi nilai variabel target, berdasarkan nilai prediktor.

3) *Prediction*

Yaitu teknik yang Menunjukkan sesuatu yang belum terjadi (terjadi di masa depan). Prediksi mirip dengan klasifikasi dan estimasi tetapi untuk prediksi, hasilnya terletak di masa depan.

4) *Classification*

Teknik yang menunjukkan Variabel tujuan adalah kategorikal. Klasifikasi mirip dengan estimasi tetapi variabel target adalah kategorikal bukan numerik.

5) *Clustering*

Teknik untuk Mengelompokkan record, observasi, atau kasus dalam suatu kelas yang memiliki kemiripan. Cluster adalah kumpulan record yang mirip satu sama lain, dan tidak mirip dengan record di cluster lain. Clustering berbeda dari klasifikasi karena tidak ada variabel target untuk clustering. Tugas pengelompokan tidak mencoba untuk mengklasifikasikan, memperkirakan, atau memprediksi nilai variabel target.

6) *Association*

Teknik yang digunakan untuk Mengidentifikasi hubungan antara peristiwa yang terjadi pada satu waktu. Tugas asosiasi untuk data mining adalah tugas menemukan atribut mana yang “berjalan bersama”.

Sementara menurut (Kantardzic, 2020) menyebutkan bahwa Teknik data mining berdasarkan tugas yang dapat dilakukan untuk tujuan prediksi dan deskripsi antara lain:

1) *Classification*

Teknik ini memiliki tugas Penemuan fungsi pembelajaran prediktif yang mengklasifikasikan item data ke dalam salah satu dari beberapa kelas yang telah ditentukan sebelumnya.

2) *Regression*

Teknik ini memiliki tugas Penemuan fungsi pembelajaran prediktif yang memetakan item data ke variabel prediksi nilai nyata.

3) *Clustering*

Teknik ini memiliki tugas deskriptif umum di mana seseorang berusaha mengidentifikasi sekumpulan kategori atau klaster terbatas untuk menggambarkan data.

4) *Summarization*

Teknik ini memiliki tugas deskriptif tambahan yang melibatkan metode untuk menemukan deskripsi ringkas untuk satu set (atau subset) data.

5) *Dependency modeling*

Teknik ini memiliki tugas Menemukan model lokal yang menjelaskan dependensi signifikan antara variabel atau antara nilai fitur dalam kumpulan data atau di bagian kumpulan data.

6) *Change and deviation detection*

Merupakan Menemukan perubahan paling signifikan dalam kumpulan data.

2.3 Klasifikasi

Tugas data mining yang paling umum dapat ditemukan di hampir setiap bidang usaha seperti perbankan, pendidikan, kedokteran, hukum dan keamanan adalah klasifikasi (Larose and Larose, 2015). Klasifikasi merupakan suatu proses untuk menemukan model atau fungsi yang menguraikan atau membedakan konsep atau kelas data (Iskandar, Refisis and Ginting, 2021). Klasifikasi adalah tugas mempelajari fungsi target f yang memetakan setiap himpunan atribut x ke salah satu label kelas y yang telah ditentukan sebelumnya seperti terlihat pada gambar 2.2 (Tan *et al.*, 2014):

Gambar 2.2 tugas memetakan atribut input x ke dalam label kelasnya

Salah satu kegunaan dari model klasifikasi adalah dapat digunakan untuk tujuan memprediksi label kelas dari catatan yang tidak diketahui (Tan *et al.*, 2014). Untuk mencapai tujuan tersebut, proses klasifikasi membentuk suatu model yang mampu membedakan data ke dalam kelas-kelas yang berbeda berdasarkan aturan atau fungsi tertentu berupa pohon keputusan, atau formula matematis (Iskandar, Refisis and Ginting, 2021). Teknik klasifikasi atau pengklasifikasi adalah pendekatan sistematis untuk membangun model klasifikasi dari kumpulan data input.

Pengklasifikasi menghasilkan model klasifikasi berdasarkan data pelatihan berisi objek yang dijelaskan oleh nilai yang mereka miliki pada sekumpulan atribut, satu atribut dibedakan sebagai kelas (Asif *et al.*, 2017).

Model yang dihasilkan harus cocok dengan data pelatihan dan memprediksi kelas atau label data yang tidak diketahui yaitu data uji yang merupakan kumpulan data terpisah yang tidak digunakan untuk menghasilkan pengklasifikasi (Asif *et al.*, 2017). Ada variabel kategori target dalam kasifikasi yang dipartisipasi ke dalam kelas atau kategori yang telah ditentukan. Jika klasifikasi dianggap dapat diterima, maka aturan yang diverifikasi dapat diterapkan pada data baru (Ramaswami *et al.*, 2019).

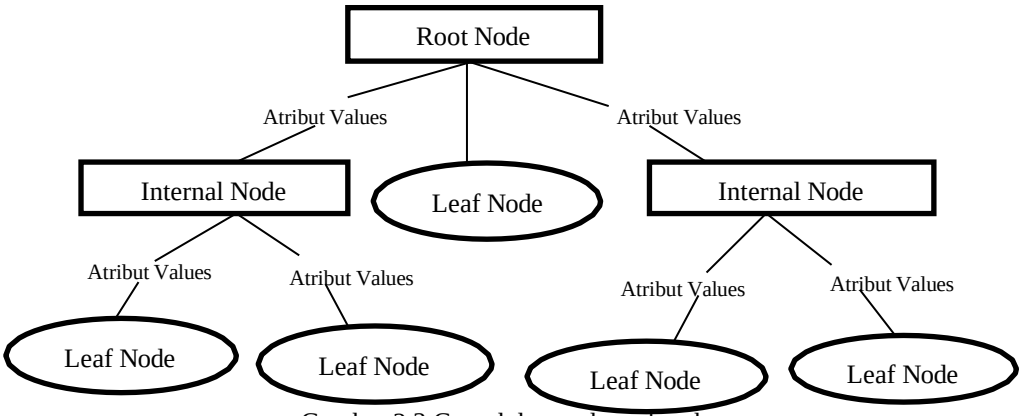
2.4 Decision Tree

Decision tree merupakan pohon keputusan classifier yang mengklasifikasikan data ke dalam label kelas yang telah ditentukan (Anggraini, Defit and Nurcahyo, 2018). Pohon keputusan atau pohon klasifikasi merupakan salah satu teknik penambangan data yang paling intuitif dan sering digunakan untuk memisahkan kumpulan data ke dalam kelas-kelas milik variabel respons (Chauhan and Kaur, 2013). Tujuannya adalah untuk membuat model klasifikasi atau dikenal sebagai classifier yang akan memprediksi dengan nilai atribut input yang tersedia. Inti dari pohon keputusan mencakup simpul akar tunggal, beberapa simpul internal dan beberapa simpul daun dan Setiap simpul daun memegang label kelas kemudian Jalur dari simpul akar ke simpul daun mengungkapkan aturan klasifikasi (Meng *et al.*, 2020).

Model pohon keputusan mengambil bentuk diagram alur keputusan atau pohon terbalik di mana atribut diuji di setiap node. Pohon keputusan memiliki 3 jenis Node (Tan, Steinbach and Kumar, 2013) yaitu :

a. *Root Node* atau simpul akar yang memiliki input cabang yang masuk dan memiliki cabang lebih dari satu, terkadang tidak memiliki cabang sama sekali yang biasanya merupakan atribut yang paling memiliki pengaruh terbesar pada suatu kelas tertentu.
b. *Interval Node* atau simpul interval yang hanya memiliki satu cabang yang masuk dan dapat memiliki lebih dari satu cabang yang keluar.
c. *Leaf Node* atau simpul daun yang merupakan simpul akhir dan hanya memiliki satu cabang yang masuk, tidak memiliki cabang sama sekali dan menandakan bahwa simpul tersebut merupakan label kelas.

Berikut contoh pohon keputusan seperti pada gambar 2.3:



Gambar 2.3 Contoh kerangka pohon keputusan

Persyaratan yang harus dipenuhi sebelum algoritma pohon keputusan dapat diterapkan (Larose and Larose, 2015) yaitu:

a. Algoritme pohon keputusan merepresentasikan pembelajaran yang diawasi sehingga membutuhkan variabel target yang telah diklasifikasi sebelumnya.

b. Kumpulan data pelatihan ini harus kaya dan beragam, memberikan algoritme penampang melintang yang sehat dari jenis rekaman yang klasifikasinya mungkin diperlukan di masa mendatang.

c. Kelas atribut target harus diskrit.

2.4.1 Algoritma C4.5

Algoritma C4.5 merupakan pengembangan dari algoritma ID3 yang menerapkan gain ratio daripada information gain sebagai standar pemilihan atribut (Meng *et al.*, 2020). Algoritma C4.5 merupakan kelompok algoritma *Decision Tree* (Iskandar, Refisis and Ginting, 2021). Bagian terpenting dari algoritma C4.5 adalah proses menghasilkan pohon keputusan awal dari kumpulan sampel pelatihan. Hasilnya, algoritma menghasilkan pengklasifikasi dalam bentuk pohon keputusan; struktur dengan dua tipe simpul daun yang menunjukkan kelas atau simpul keputusan yang menentukan beberapa pengujian yang akan dilakukan pada nilai atribut tunggal, dengan satu cabang dan subpohon untuk setiap kemungkinan hasil pengujian (Kantardics, 2020)

Tahap untuk menghasilkan pohon keputusan menggunakan algoritma C4.5 sebagai berikut (Hana, 2020) :

1. Membuat pohon keputusan dengan membentuk akar pohon kemudian dibedakan sesuai atribut yang serasi untuk daun
2. *Tree pruning* yaitu proses pemangkasan cabang pohon yang tidak diperlukan oleh pohon yang sudah terbentuk atau dengan kata lagi dilakukan penyederhanaan ukuran pohon karena pohon keputusan yang dibentuk biasanya bentuknya besar. Alasan lain pruning harus dilakukan adalah karena

dalam teknik klasifikasi yang akan dijalankan nantinya akan mengeluarkan rule (pola) yang dibentuk berdasarkan struktur tree, jadi jika struktur tree tidak teratur atau kurang sederhana, maka *rule* yang dihasilkan pun akan rumit untuk diimplementasikan (Khairul Amin and Sibaroni, 2015).

3. Pembuatan aturan keputusan dengan tindakan penelusuran dari akar sampai ke daun, aturan (*rule*) tersebut diturunkan dari pohon keputusan.

Algoritma C4.5 tidak hanya dapat memproses data diskrit tetapi juga dapat memproses data dengan atribut kontinu (Han, Ma and Yang, 2018). Sementara langkah untuk membangun pohon keputusan adalah sebagai berikut (Anggraini, Defit and Nurcahyo, 2018) :

a) Pilih sebuah atribut sebagai *node/akar*

Dasar untuk memilih atribut sebagai akar adalah nilai gain tertinggi dari atribut-atribut yang ada. Nilai Gain merupakan tingkat pengaruh suatu atribut terhadap keputusan atau ukuran efektivitas suatu variabel dalam mengklasifikasikan data (Setio, Saputro and Bowo Winarno, 2020). Untuk mendapatkan nilai Gain maka harus menghitung nilai Entropy terlebih dahulu untuk mengukur heterogenitas suatu kumpulan data (Khairul Amin and Sibaroni, 2015). Semakin kecil nilai entropy, semakin baik digunakan dalam mengekstraksi suatu kelas (Setio, Saputro and Bowo Winarno, 2020). Secara matematis nilai entropy dapat dihitung dengan menggunakan formula seperti yang tertera dalam persamaan 2.2 sebagai berikut: :

$$Entropy\ (S) = -\sum_{i=1}^{n} Pi * log_2 Pi \quad \text{..... (2.1)}$$

Keterangan :

S = Himpunan kasus

n = Jumlah partisi S

Pi = Proporsi dari Si terhadap S

Langkah selanjutnya adalah menghitung nilai Gain dari suatu atribut. Adapun formula untuk menghitung nilai gain dapat dilihat pada persamaan 2.2 sebagai berikut:

$$Gain\ (S, A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{S} * Entropy\ (Si) \dots \dots \dots (2.2)$$

Keterangan :

S = Himpunan Kasus

A = Atribut

n = Jumlah partisipasi atribut A

$|Si|$ = Jumlah kasus pada partisipasi ke i

$|S|$ = Jumlah kasus dalam S

Meskipun kriteria gain memiliki beberapa hasil yang baik dalam konstruksi pohon keputusan, tetapi kriteria ini juga memiliki satu kekurangan yang serius yaitu bias yang kuat dalam mendukung pengujian dengan banyaknya informasi. Oleh krena itu sebuah solusi ditemukan dalam beberapa jenis normalisasi dengan parameter tambahan untuk menghindari kelebihan informasi dan menormalkan perolehan informasi yaitu dengan menghitung split info dan gain ratio (Kantardics, 2020).

Tujuan dari perhitungan Information Gain dan Split Info adalah untuk mendapatkan nilai Gain Ratio. Kriteria gain ratio memilih tes yang

memaksimalkan rasio sebelumnya untuk memberikan pilihan tes yang lebih baik dan konsisten daripada kriteria perolehan sebelumnya (Kantardics, 2020). Berikut formula mencari nilai split info dan gain ratio seperti pada persamaan 2.3 dan persamaan 2.4:

Persamaan Split Info :

$$Split\ Info\ (S,A) = -\sum_{j=1}^{k} \frac{S_i}{S} x log_2 \frac{S_i}{S} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2.3)$$

Keterangan :

S = Data sampel yang digunakan untuk data training

A = Atribut

n = Jumlah partisipasi atribut A

$|Si|$ = Jumlah sampel untuk atribut i

$|S|$ = Jumlah kasus dalam S

Persamaan Gain Ratio :

$$Gain\ Ratio\ (S,A) = \frac{Gain\ (S,A)}{Split\ (S,A)} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2.4)$$

b) Membuat cabang untuk setiap nilai.

Pemilihan root/cabang dilakukan berdasarkan gain ratio terbesar setelah menghapus atribut yang telah terpilih sebagai root (Khairul Amin and Sibaroni, 2015).

c) Membagi kasus dalam cabang.

d) Ulangi proses untuk masing-masing cabang sampai semua kasus pada cabang memiliki kelas yang sama. Proses percabangan akan berhenti apabila semua kasus dalam simpul n mendapat kelas yang sama, tidak ada variabel

independen di dalam kasus yang dipartisi lagi, tidak ada kasus di dalam cabang yang kosong (Setio, Saputro and Bowo Winarno, 2020).

2.4.2 Algoritma Random Forest

Random forest merupakan pengembangan dari metode ensemble yang pertama kali dikembangkan oleh Leo Breman tahun 2001, dimana metode ini digunakan untuk meningkatkan ketepatan klasifikasi dengan membentuk gabungan pohon klasifikasi (CART) yang saling independen dan klasifikasi diperoleh melalui proses voting dengan mengambil jumlah terbanyak dari pohon-pohon keputusan yang terbentuk (Anouze and Bou-Hamad, 2019). Semakin banyak pohon keputusan yang dihasilkan algoritma, maka semakin akurat hasilnya (Nachouki and Naaj, 2022). Model ini membuat pohon keputusan yang berbeda berdasarkan sampel data dan ketika poin data baru dimasukkan untuk prediksi kelasnya, setiap pohon keputusan memberikan satu prediksi, dan akhirnya solusi terbaik dipilih dengan pemungutan suara (Ghosh and Janan, 2021).

Kelebihan dari Random Forest antara lain mampu menghasilkan kesalahan yang lebih rendah, memberikan hasil klasifikasi yang baik, mampu menangani data pelatihan dalam jumlah yang sangat besar secara efisien, dan metode yang efektif untuk memperkirakan data yang hilang (Breiman, 2001). Konsep Random Forest adalah unik karena dapat digunakan untuk menangani data dalam jumlah yang besar dan mudah untuk menangani data yang hilang. Sementara secara statistik, Random Forest menarik karena tersedianya fitur tambahan seperti ukuran kepentingan variabel, pembobotan kelas diferensial, imputasi nilai yang hilang dan visualisasi (Cutler, Cutler and Stevens, 2012).

Dari sudut pandang komputasi, Random Forest menarik karena beberapa alasan diantaranya (Cutler, Cutler and Stevens, 2012):

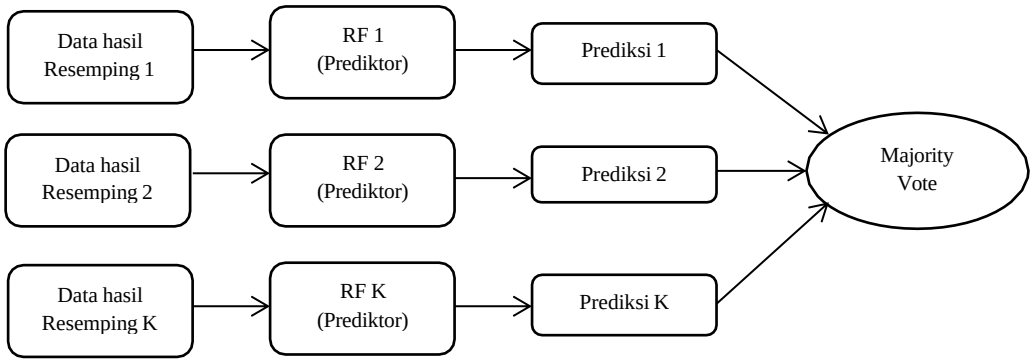
a. Secara alami menangani klasifikasi regresi dan (multikelas);
b. Relatif cepat untuk dilatih dan diprediksi;
c. Hanya bergantung pada satu atau dua parameter penyetelan;
d. Memiliki estimasi kesalahan generalisasi bawaan;
e. Dapat digunakan langsung untuk masalah dimensi tinggi;
f. Dapat dengan mudah diimplementasikan secara paralel.

Adapun langkah pembentukan pohn keputusan dengan Random Forest dapat dijelaskan sebagai berikut (Fachruddin, 2015):

a. Mengambil n data sampel dari dataset awal dengan menggunakan teknik resempling bootstrap dengan pengembalian
b. Menyusun pohon klasifikasi dari setiap dataset hasil resempling bootstrapdengan menentukan pemilah terbaik yangdidasarkan pada variabel prediktor yang diambil secara acak dengan perhitungan log2 (M+1), dimana M adalah banyak variabel prediktor (Breiman, 2001) atau $\frac{1}{2}|\sqrt{p}|$ $|\sqrt{p}|$, $2|\sqrt{p}|$,dimana p adalah banyak variabel prediktor (Lumbanraja *et al.*, 2019), (Pangastuti, 2018).
c. Melakukan prediksi klasifikasi data sampel berdasarkan pohon klasifikasi yang terbentuk.
d. Mengulangi langkah pada poin a sampai c sebanyak k kali sehingga diperoleh pohon klasifikasi yang diinginkan

e. Melakukan prediksi klasifikasi data sampel akhir dengan mengombinasikan hasil prediksi pohon klasifikasi yang diperoleh berdasarkan aturan majority vote.

Berikut contoh pohon keputusan dari Random Forest seperti pada gambar 2.4:



Gambar 2.4 Ilustrasi Random forest (Fachruddin, 2015)

Ada banyak metode pemilihan atribut tetapi ukuran pemilihan atribut yang paling sering digunakan dalam induksi pohon keputusan adalah Gain dan Indeks Gini. Tetapi Random Forest Classifier menggunakan metode Indeks Gini (Ghosh and Janan, 2021). Indeks Gini digunakan untuk memilih fitur di setiap simpul internal dari pohon keputusan (Suci Amaliah, Nusrang and Aswi, 2022). Berikut formula mencari nilai Indeks Gini seperti pada persamaan sebagai berikut:

$$Gini\ (S_i) = 1 - \sum_{i=0}^{c-1} P_i^2 \quad \cdots\cdots (2.5)$$

Keterangan :

P_i = Frekuensi relatif kelas Ci di dalam set

Ci = Kelas untuk i = 1

C-i dan C = Jumlah kelas yang telah ditentukan

Sementara kulaitas split pada fitur *k* ke dalam subset S_i merupakan jumlah sampel milik kelas C_i, kemudian dihitung sebagai jumlah perimbangan indikasi Gini dari subset yang dihasilkan. Adapun data dapat dihitung dengan rumus sebagai berikut (Suci Amaliah, Nusrang and Aswi, 2022):

$$Gini_{split} = \sum_{i=0}^{k=1}\left(\frac{n_i}{n}\right) Gini \quad (S_i) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.6)$$

Keterangan :

n_i = Jumlah sampel dalam subset S_i setelah di split

n = Jumlah sampel di node yang diberikan

2.5 Evaluasi metode klasifikasi

Evaluasi kinerja model klasifikasi didasarkan pada jumlah catatan pengujian yang diprediksi dengan benar dan salah oleh model yang dikenal sebagai matriks konfusi. matriks konfusi memberikan informasi yang diperlukan untuk menentukan seberapa baik kinerja model klasifikasi, meringkas informasi ini dengan satu nomor akan membuatnya lebih mudah untuk membandingkan kinerja model yang berbeda (Tan *et al.*, 2014).

2.5.1 K-Ford Validation

Cross Validation adalah salah satu metode resampling data yang paling banyak digunakan untuk memperkirakan kesalahan prediksi sebenarnya dari model (Berrar, 2018). Dalam k-fold cross-validation, set pembelajaran yang tersedia dipartisi menjadi k subset terpisah dengan ukuran yang sama. Proses training dan testing dilakukan sebanyak k kali. Dalam iterasi ke-i, data partisi Di

diposisikan sebagai data testing, sementara partisi lain yang tersisa secara kolektif digunakan untuk melatih model (Pangastuti, 2018).

Model dalam klasifikasi dilatih dan diuji sebanyak K. Disetiap pengulangan, salah satu himpunan bagian akan digunakan sebagai data training dan data testing. Adapun langkah-langkah dari k fold cross validation adalah sebagai berikut:

1. Total data dibagi menjadi k bagian.
2. Fold ke-1 adalah ketika bagian ke-1 menjadi data uji (*testing data*) dan sisanya menjadi data latih (*training data*). kemudian, hitung akurasi atau kesamaan atau kedekatan suatu hasil pengukuran dengan angka atau data yang sebenarnya berdasarkan porsi data tersebut.
3. Fold ke-2 adalah ketika bagian ke-2 menjadi data uji (*testing data*) dan sisanya menjadi data latih (*training data*). Selanjutnya, hitung akurasi berdasarkan porsi data tersebut.
4. Demikian seterusnya hingga mencapai fold ke-k. Hitung rata-rata akurasi dari k buah akurasi diatas. Rata-rata akurasi ini menjadi akurasi final.

Pada dasarnya tidak ada ketentuan dalam membagi Nilai k, tetapi dengan Nilai k yang terlalu kecil akan menghasilkan model yang serupa dengan metode cross validation biasa yang hanya membagi data menjadi latih-tes saja dan dapat memicu terjadinya bias. Sementara dengan nilai k yang terlalu besar akan menghasilkan model yang tidak bias tetapi dapat membuat variansi menjadi besar sehingga dapat memicu terjadinya overfit (Widyaningsih, Arum and Prawira, 2021).

Untuk mengurangi evaluasi yang bias dan overfit, direkomendasikan menggunakan nilai k yang sedang yaitu 10-20-fold cross-validation (Berrar, 2018). 10-fold cross-validation merupakan cara standar untuk memprediksi tingkat kesalahan dari teknik pembelajaran yang diberikan satu sampel data tetap tetapi 5 -fold cross-validation atau 20 -fold cross-validation kemungkinan besar hampir sama baiknya seperti 10-fold cross-validation (Witten, Frank and Hall, 2017).

2.6 Penelitian terkait

Data mining telah diterapkan dalam lingkungan pendidikan untuk memprediksi kinerja akademik mahasiswa. Berikut beberapa penelitian penerapan data mining yang dterapkan untuk memprediksi kinerja akademik mahasiswa seperti yang terlihat pada tabel 2.1:

Tabel 2.1 Penelitian terkait

No	Judul, penulis, tahun	Metode	Hasil
1	Perbandingan Algoritma C4.5 Dan Id3 Untuk Prediksi Ketepatan Waktu Lulus Mahasiswa (Faizah and Jananto, 2021)	Algoritma C4.5 Dan Id3	Hasil pengujian menunjukkan bahwa Secara keseluruhan hasil implementasi terhadap komposisi data 90%, 80%, 70%, 30%, 20%, dan 10% diperoleh informasi bahwa nilai akurasi yang dihasilkan

			menggunakan algoritma C4.5 rata-rata lebih tinggi dibandingkan dengan nilai akurasi algoritma ID3. Sehingga dapat disimpulkan algoritma terbaik untuk memprediksi ketepatan kelulusan mahasiswa adalah algoritma C45.
2	Komparasi Algoritma *Decision Tree, Naive Bayes* Dan *K-Nearest Neighbor* Untuk Memprediksi Mahasiswa Lulus Tepat Waktu (Budiyantara *et al.*, 2020)	Decision Tree, Naive Bayes Dan K-Nearest Neighbor	Hasil evaluasi dan validasi menunjukkan bahwa hasil akurasi dari Metode *Decision Tree* (C4.5) sebesar 98.04% pada pengujian ke 3. akurasi Metode *Naïve Bayes* sebesar 96.00% pada pengujian ke 4. Dan akurasi Metode *K-Nearest Neighbor* (K-NN) sebesar 90.00% pada pengujian ke 2.
3	Comparative Study of	C5.0, J48,	hasil Akurasi Random

	Supervised Algorithms for Prediction of Students' Performance (Sathe and Adamuthe, 2021)	CART, Naïve Bayes (NB), K-Nearest Neighbour (KNN), Random Forest and Support Vector	Forest menunjukkan sebesar 98.12%
4	Student Career Prediction Using Decision Tree and Random Forest Machine Learning Classifiers (VidyaShreeram and Muthukumaravel, 2021)	DT dan RF	menghasilkan akurasi Decision Tree sekitar 91% dan tingkat akurasi Random Forest sekitar 93%
5	Machine Learning Approaches to Predict Learning Outcomes in Massive Open Online Courses (Al-Shabandar *et al.*, 2017).	Logistic Regression (LR), Linear Discriminant Analysis (LDA), Naive Bayes (NB), Support Vector Machine	hasil menunjukkan bahwa Random Forest kinerja tertinggi dari model yang diuji.yaitu diperoleh nilai area Under Curve (AUC) sebesar 0,90

		(SVM), Decision Tree (DT), Random Forest (RF), Neural Network (MLP) dan Self Organized Map (SOM)	
6	Using educational data mining techniques to increase the prediction accuracy of student academic performance (Ramaswami *et al.*, 2019).	Naïve Bayes, Random Forest, Statistik Logistik dan k-NN	hasil menunjukkan bahwa Random Forest lebih akurat daripada tiga algoritma lainnya dengan cara yang signifikan secara statistik yaitu 88%

Berdasarkan beberapa hasil *review* penelitian baik jurnal nasional dan internasional terkait yang telah disebutkan pada pada tabel 2.1 diatas dapat disimpulkan bahwa dari beberapa metode yang digunakan pada jurnal yang peneliti *review* tingkat akurasi cenderung lebih tinggi dengan menggunakan algororitma C4.5 dan Algoritma Random Forest dengan akurasi tertinggi mencapai 98.04% dan 98,12 %. Hasil kesimpulan yang diperoleh dari hasil *review* penelitian terdahulu menjadi dasar atau alasan penulis untuk menggunakan teknik

data mining dengan algororitma C4.5 dan Random Forest untuk melakukan penelitian di bidang data mining pendidikan dalam memprediksi kinerja akademik mahasiswa khususnya berkaitan dengan masa studi.